Generating math learning materials with LLMs

Overview

01

One-sentence focus

Exploring how NLP and Large Language Models can assist in creating accurate and relevant learning materials for mathematics education.

02

Keywords

Education, Studying, Coursework, Preparation, Generation

03

Research Question & Motivation

Current AI tools tend to deviate from specific course topics when asked to generate material. There’s a need for reliable and accurate AI-generated educational materials. This would have the potential to support teachers and students in personalized learning

Methodology & Tools

Tools

Ollama
Library allowing for easy model usage and downloading

Python
Used to help build our application and allow users to access the different models and tools we provide

PyTorch
Ease of training a new model for MWP tagging based on the MWP-BERT design

Model Capabilities

Marked Up Exam Labeling
Our application will have different models that can label exam topics, give answers, and grade user answers

Question Generation
Potentially use another model to compare if question generated matches topics

Grading
Give model answers to question and compare output grade to human generated grade from datasets

Testing

Answer Checking
Have model generate answers and compare to human answers to same questions

Datasets
Google's mathematics dataset - QA pairs
Math23k - MWP dataset
GSM8K - US state exams

Evaluation
Compare grade, topic, and answer accuracy in materials generated and processed

Testing GPT-oss Accuracy → Research into how accurate GPT-oss is at answering different types of math questions

- How often does the model get questions of a certain type wrong.
- Does the model get the same questions wrong repeatedly.
    - If so why?
    - What would this entail for the question?
- What types of questions does it struggle with most.
- Could it be used to generate accurate answers to questions for a test.

GPT-oss Answer Accuracy

GPT-oss Accuracy

→

For all questions the model was prompts to show all step for solving and then provide a final answer which was compared against the true correct answer.

Word problems.
Example: James decides to bulk up. He weighs 120 kg and gains 20% of his body weight in muscle and 1 quarter that much in fat. How much does he weigh now?
Accuracy: 94.961%
Accuracy when prompted to forgo reasoning: 95.259% (and faster)

Calculus Derivatives.
Example: Find the third derivative of 178776b^4 + 312b^3 - 3594b^2 + 68923b with respect to b.
Accuracy: 97.714%

Algebra: Linear Systems
Example: Let i = -1285 + 1287. Solve -102*w + 100*w - 2*p + i = 0, 3*p + 1 = -w for w.
Accuracy: 100% (less runs around 300)

Why are word problems the least accurate?

Repeated Errors

→

When the model answered the first 300 word problems over and over. Half of the mistakes it made were always on the same questions.

Jasper will serve charcuterie at his dinner party. He buys 2 pounds of cheddar cheese for $10, a pound of cream cheese that cost half the price of the cheddar cheese, and a pack of cold cuts that cost twice the price of the cheddar cheese. How much does he spend on the ingredients?

If the cheddar cheese costs \$10 for 2 lb, the price per pound of cheddar is
\$10 ÷ 2 lb = <<10/2=5>>5 dollars per pound.
The cream cheese costs half the price of the cheddar cheese (per pound).
Half of \$5 is
\$5 × ½ = <<5/2=2.5>>2.5 dollars for the 1 lb of cream cheese.
The cold cuts cost twice the price of the cheddar cheese (the total price of the cheddar).
Twice \$10 is
\$10 × 2 = <<10*2=20>>20 dollars.
Now add all three amounts:
Cheddar: \$10
Cream cheese: \$2.5
Cold cuts: \$20
Total spent = \$10 + \$2.5 + \$20 = <<10+2.5+20=32.5>>32.5 dollars.

What repeated errors mean.

Repeated Errors

→ Impact of repeated error on evaluation

The question is vague.
The previous question could be interpreted in multiple ways that could be considered correct.

Repetitions impact on accuracy.
Repetition would only improve accuracy but not guarantee it.
It would weed out bad answers but not bad questions.
It corrects for model errors in reasoning not with interpretation of questions.
Improves accuracy for complex non word problems.

Could help recognize a bad question.
If you have the answers to a question and run it thought the model and it gets it wrong almost every time,There may be some vagueness or uncertainty to the question.

Deepseek-r1 topic generation

Question Generation →

Although slow, questions were generated with higher accuracy of course material the simpler the subject matter was

- First pass was through human grading of accuracy
  - Of 50 questions generated for a variety of math topics, it was able to stick to requested parameters 92% of the time
  - The times it didn't stick on topic it just included slightly more advanced parts of a problem than would be expected to be solvable for a grade level
    - EX: Integrals/Riemann Sums when in early pre-calculus
- Second pass was with another model running through questions generated
- In the process of labeling math dataset questions to see accuracy
  - Google Deepmind's Mathematics Dataset
  - Math23k

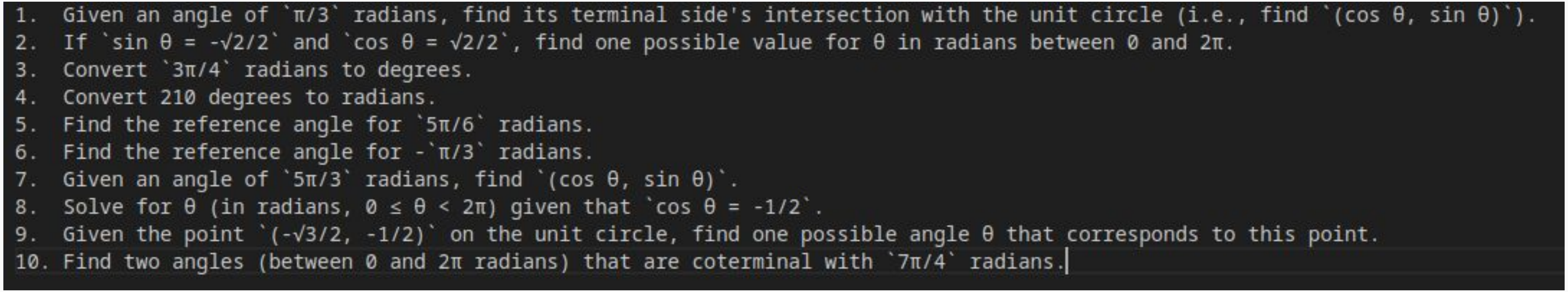
Deepseek-r1 topic generation

System Prompt

"Write 10 math problems for the given math subject for the given grade level."

User Prompt

"Trigonometry problems about the unit circle and finding angles from radians for 11th grade students"



```
1.  Given an angle of `π/3` radians, find its terminal side's intersection with the unit circle (i.e., find `(cos θ, sin θ)`).
2.  If `sin θ = -√2/2` and `cos θ = √2/2`, find one possible value for θ in radians between 0 and 2π.
3.  Convert `3π/4` radians to degrees.
4.  Convert 210 degrees to radians.
5.  Find the reference angle for `5π/6` radians.
6.  Find the reference angle for -`π/3` radians.
7.  Given an angle of `5π/3` radians, find `(cos θ, sin θ)`.
8.  Solve for θ (in radians, 0 ≤ θ < 2π) given that `cos θ = -1/2`.
9.  Given the point `(-√3/2, -1/2)` on the unit circle, find one possible angle θ that corresponds to this point.
10. Find two angles (between 0 and 2π radians) that are coterminal with `7π/4` radians.
```

Deepseek-r1 identification accuracy

Topic Labeling

Questions were fed into models and asked to give:

- Grade level
- Topic
- Related keywords to help with studying

- Deepseek performed well at labeling topics created by itself, however:
  - It frequently adjusted what grade level it thought could solve questions
    - Some calculus questions were labeled as ready for 10th graders, others for 12th graders
  - Add in keywords or topics that would not be helpful in solving problems
    - Ex: Quadrant analysis on angle→radian conversions
  - Overall in labeling it got the overall topic right 95% of the time, but 76% of the time some sort of unhelpful information was generated
    - Superfluous labels and varying grade levels were the only thing that tended to get wrong, and this appeared in almost all analysis it did
- Topics that are easier or more ingrained in required education are the easiest to label for.
  - The more googlable the subject matter, the more accurate the model was

Deepseek-r1 topic identification

System Prompt

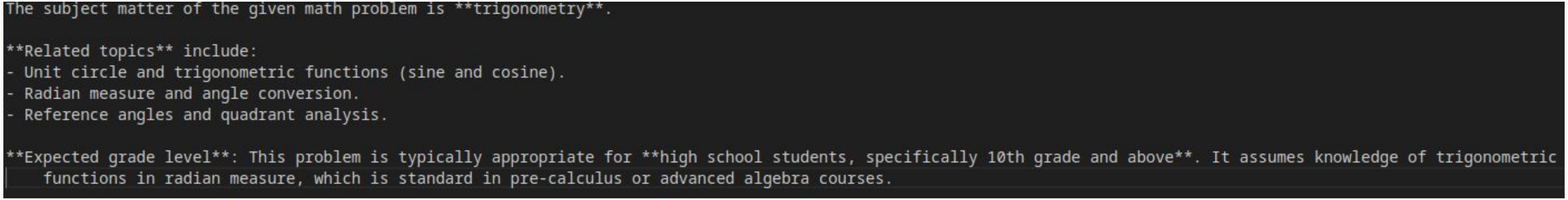
→

"Identify the subject matter of a given math problem input with related topics and expected grade level for the question."

User Prompt

→

"Find the angle for -`π/3` radians."

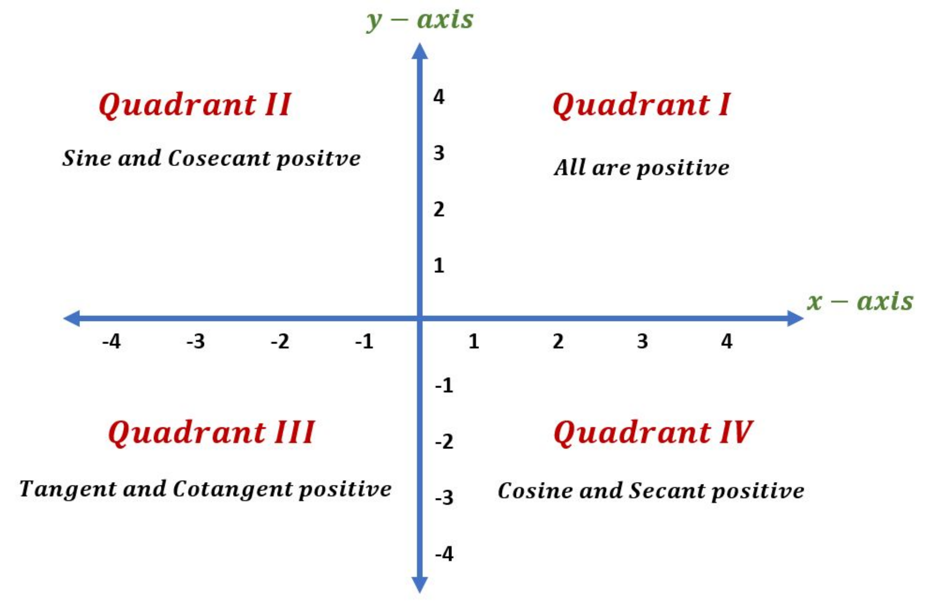


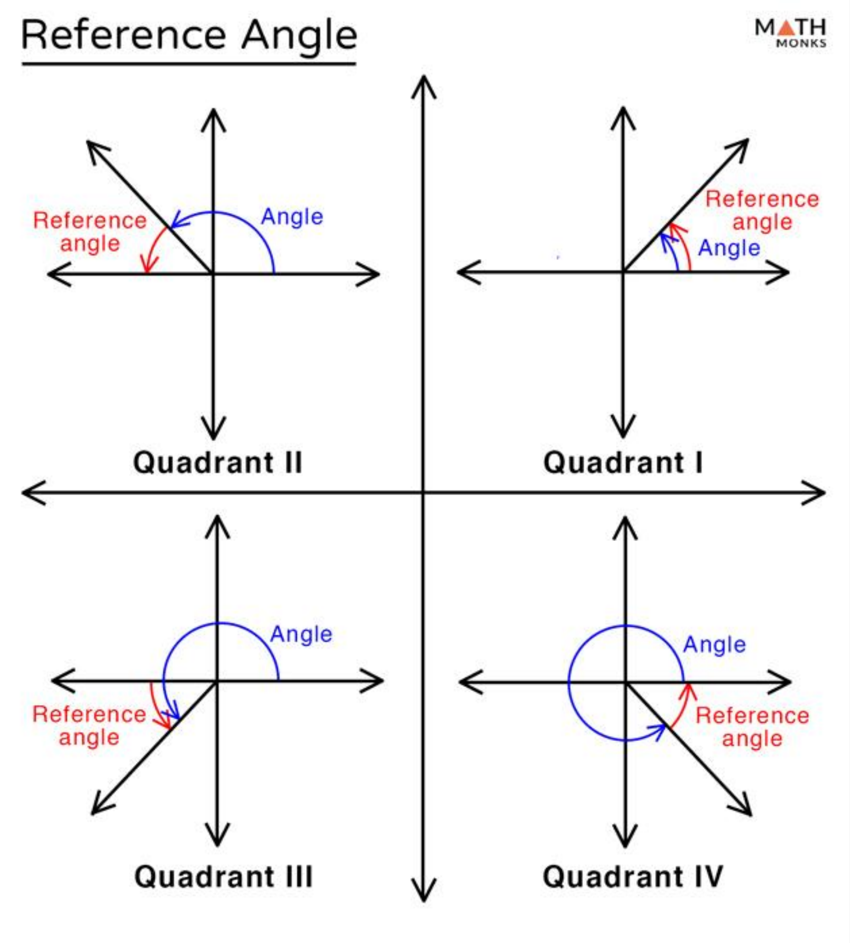
```
The subject matter of the given math problem is **trigonometry**.

**Related topics** include:
- Unit circle and trigonometric functions (sine and cosine).
- Radian measure and angle conversion.
- Reference angles and quadrant analysis.

**Expected grade level**: This problem is typically appropriate for **high school students, specifically 10th grade and above**. It assumes knowledge of trigonometric
    functions in radian measure, which is standard in pre-calculus or advanced algebra courses.
```

What the model suggests to research

$$\text{Radians} = \left(\frac{\pi}{180^0}\right) \times \text{degrees}$$

$$\text{Degrees} = \left(\frac{180^0}{\pi}\right) \times \text{radians}$$





What’s the most helpful piece of information

$$\text{Radians} = \left(\frac{\pi}{180^{0}}\right) \times \text{degrees}$$

$$\text{Degrees} = \left(\frac{180^{0}}{\pi}\right) \times \text{radians}$$

What’s next?

Problem Labeler

MWP-BERT is a POS tagger designed for word problems. This would allow us to pull out numbers from problems and place them into a calculator for better answer acccuracy

Question Generation

Models seem to perform the best when generating questions relating to more commonly taught fields of mathematics

Topic Labeling

This is mainly our test to see how accurate our questions are to what users requested. This is part of our data validation. As of right now, we need to do more work (as outlined earlier) to ensure our accuracy

Question Answering

Making sure that the answers to the generated questions are tested for accuracy. Running it through a model multiple times and comparing answers.

Questions?